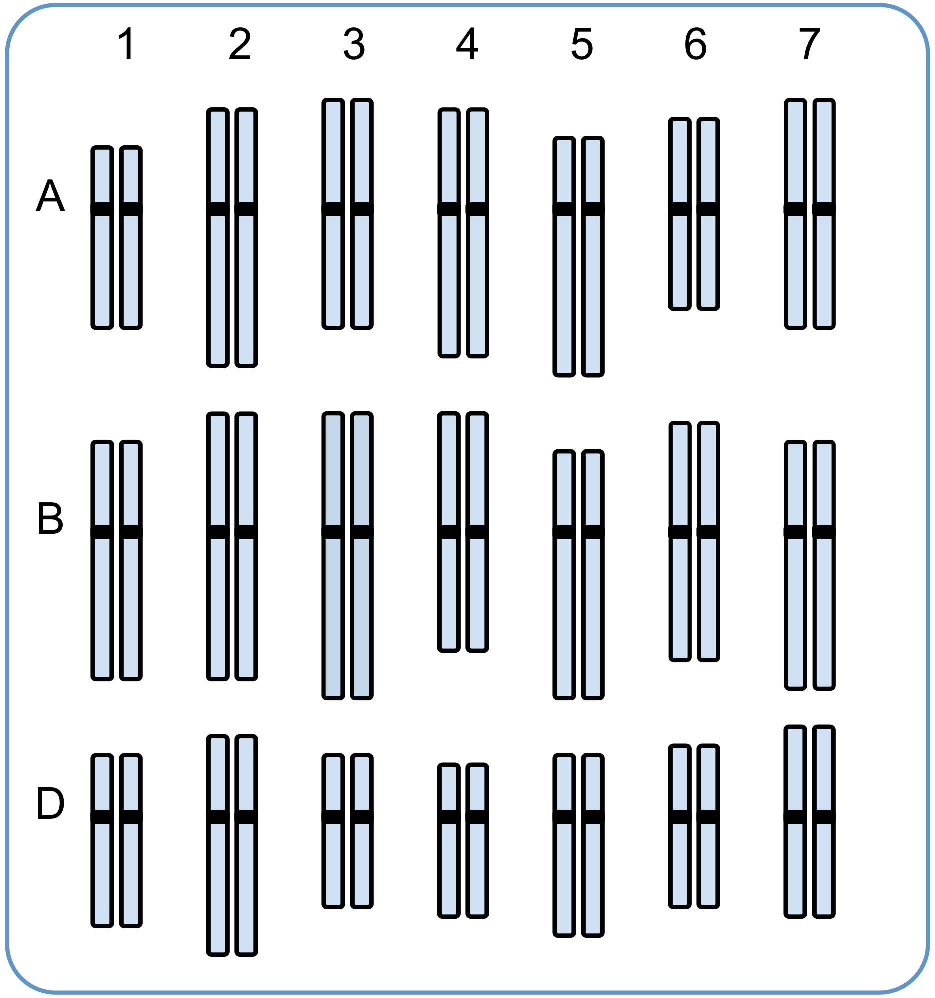
IWGSC S&P workshop, Jan 17th 2017

Challenges of the gene and repeat annotation of the wheat genome

Frédéric Choulet

GDEC, INRA, UCA, Clermont-Ferrand, France



IWGSC RefSeq v1

21 pseudomolecules

❑ TriAnnot – gene modeling

- *Hélène Rimbert*
- *Philippe Leroy*

- Uses 4 gene-modeling approaches
  - Augustus, FGENESH (ab-initio)
  - SIMSearch
  - BLASTX-Exonerate

- **Scoring**
  - based on alignment with best hit in *Poaceae* proteomes
  - score = **`QueryCoverage*2 + HitCoverage + %Identity`**
    `+/- penalties`

➔ *1 CDS per locus*

- Assign **confidence index** (of the <u>gene structure)</u>
  - **HC/LC:** agreement with mapped evidence
  - **Pseudogenes**

- **Filtering out** doubtful predictions (=discard models with no similarity with RNASeq/IsoSeq OR proteins)

- Cufflinks (RNASeq)

➔ *splicing variants + UTRs*

Proteins

maize

sorghum

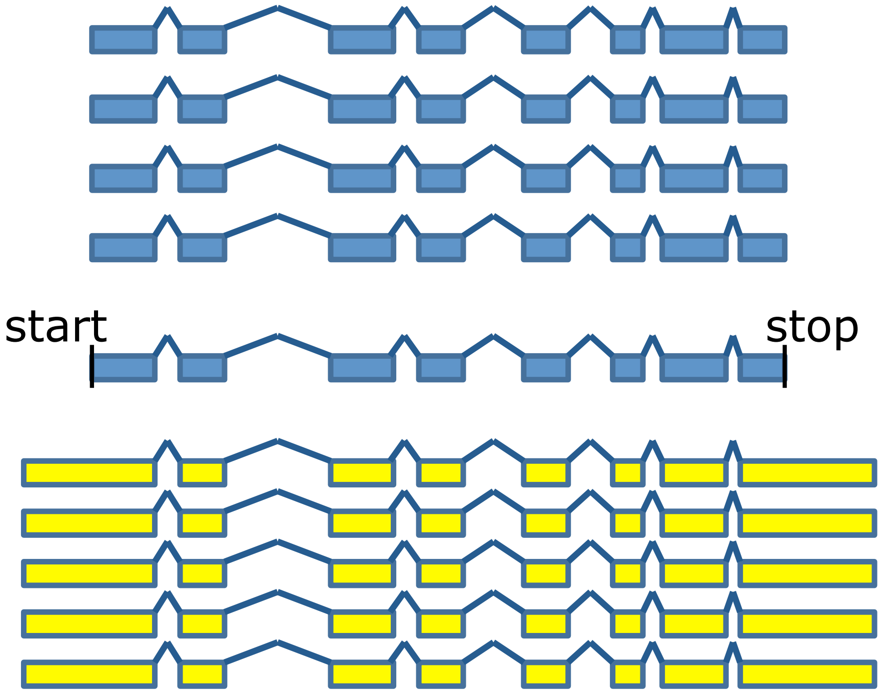
rice

brachy

CDS model

High Conf

RNASeq samples

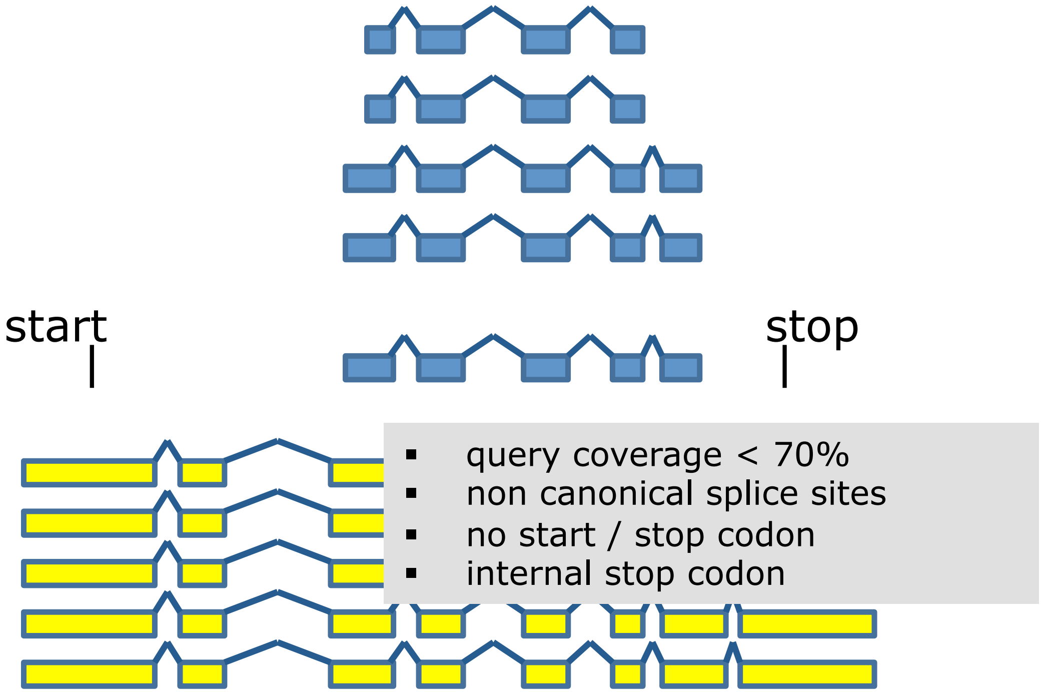


Proteins

CDS model

Pseudogene

RNASeq samples



- query coverage < 70%
- non canonical splice sites
- no start / stop codon
- internal stop codon

❑ To do

- Protein-coding
  - **Combining** 2 sets of gene predictions
    - TriAnnot
    - PGSB pipeline
  - Integrate manually curated gene families
  - Function assignment
- ncRNAs??????
- NTRs (novel transcribed regions)

PGSB, Munich

- *S. Twardziok*
- *M. Spannagl*
- *K. Mayer*

IE, Norwich

- *D. Swarbreck*
- *L. Venturini*

❑ Homeologs / Orthologs / Paralogs

➢ *Romain De Oliveira*

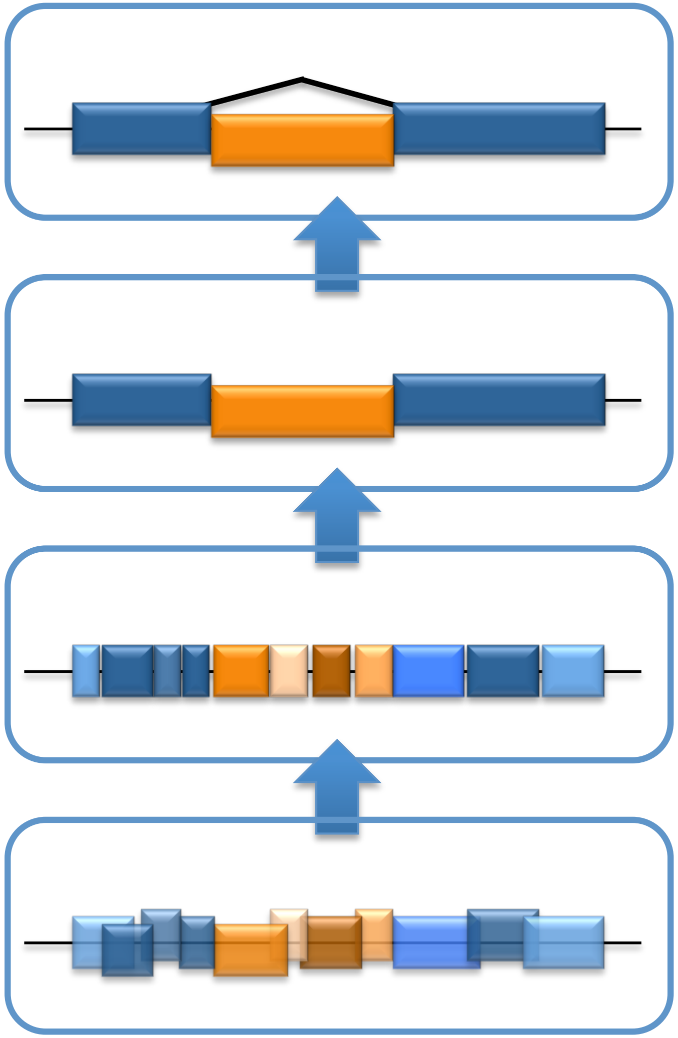
work in progress…

- 5 species:
wheat, rice, Brachypodium, maize, sorghum

- 3 tested approaches:
  - orthoMCL *(Li et al. Genom Res 2003)*
  - Silix *(Miele et al. BMC Bioinfo 2011)*
  - OMA *(Altenhoff et al. NAR 2015)*

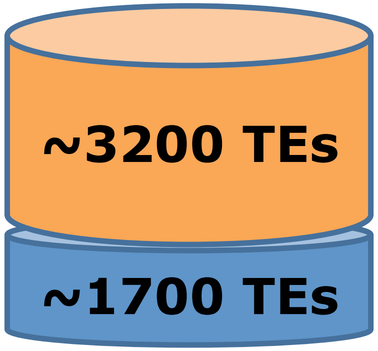
=> define **homeologs, orthologs, gene families**

❑ TE modeling with CLARI-TE



Daron et al. Genome Biol 2015

❑ ClariTeRep



~3200 TEs

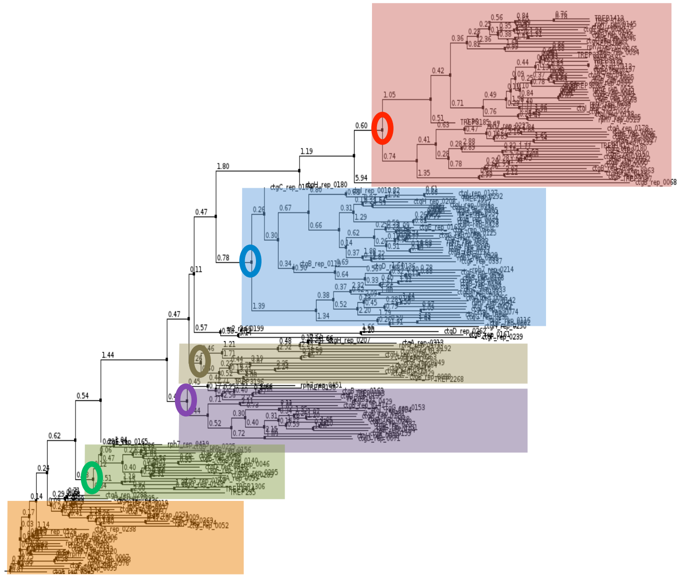
~1700 TEs

Choulet et al. 2010

CACTA++

TREP

=> **MCL** denovo clustering + manual curation



RLG_famc1.1

RLG_famc1.2

RLG_famc1.3

RLG_famc1.4

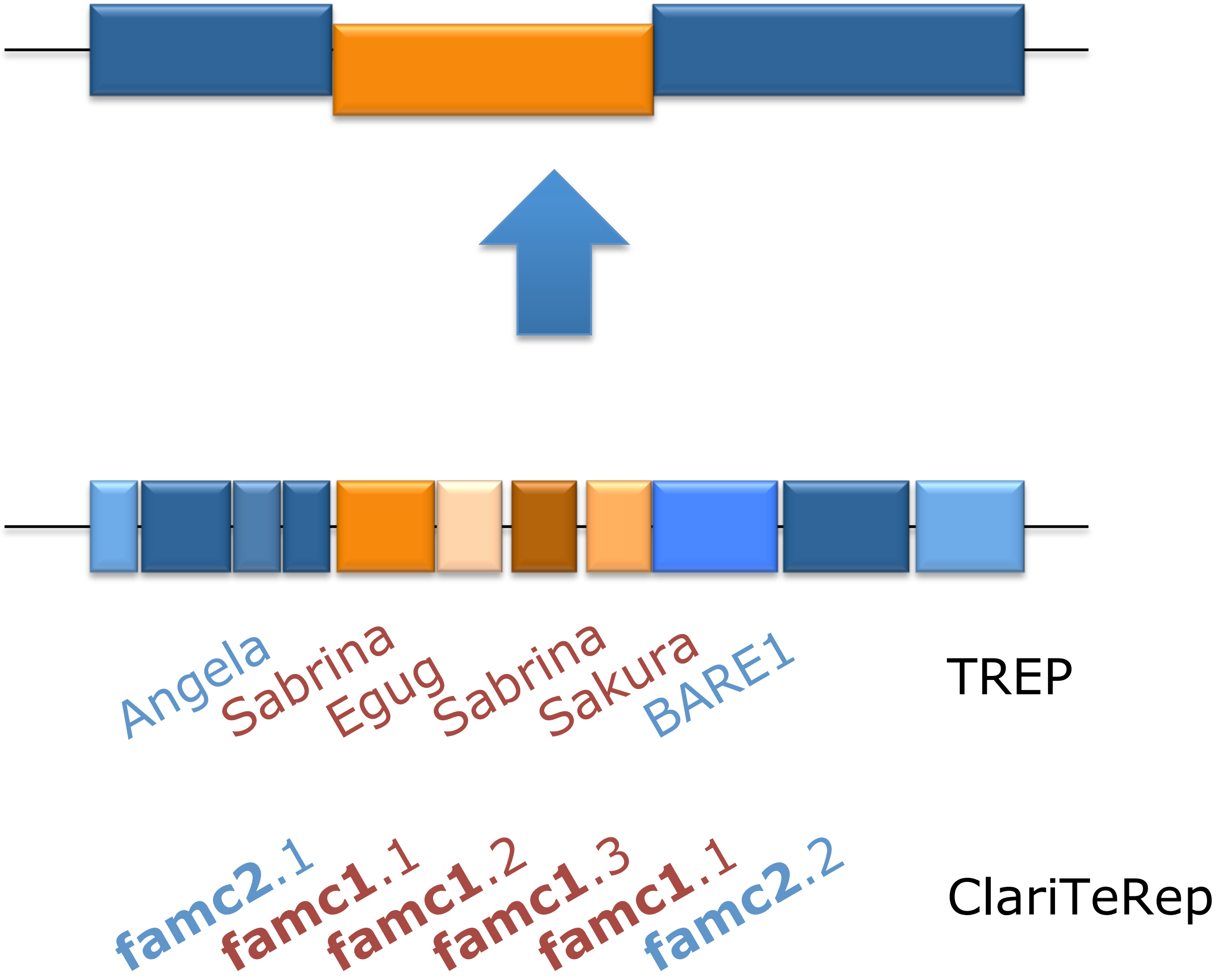
RLG_famc1.5

ClariTeRep

~500 classified families

download @ github.com/…/CLARI-TE

- CLARI-TE uses new classification for defragmentation



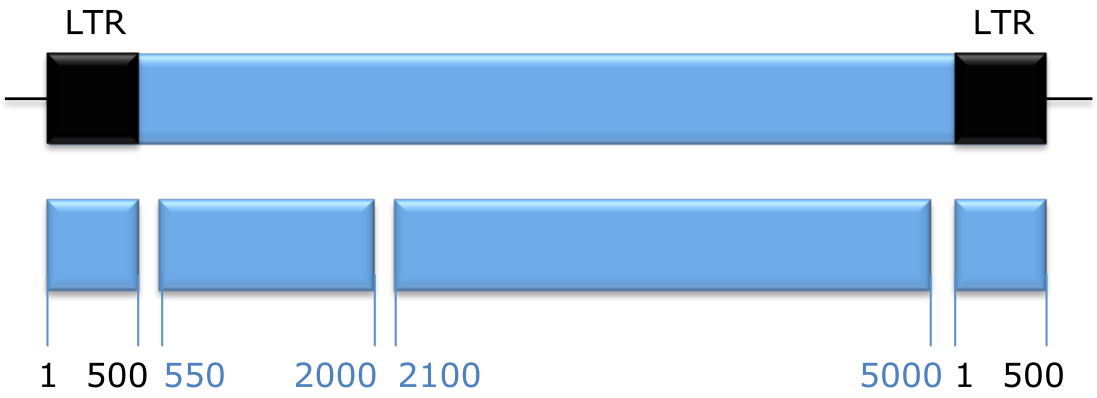
Angela Sabrina Egug Sabrina Sakura BARE1

TREP

famc2.1 famc1.1 famc1.2 famc1.3 famc1.1 famc2.2

ClariTeRep

- CLARI-TE uses new classification for better defragmentation
- CLARI-TE uses info from LTR coordinates

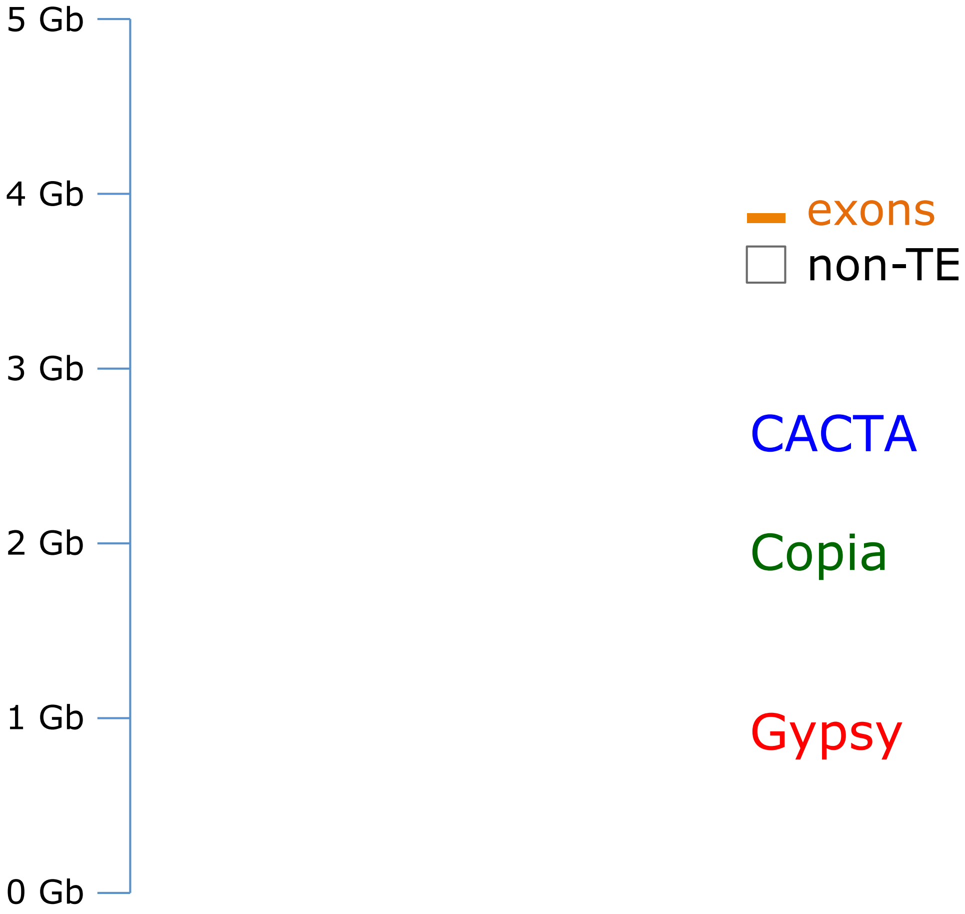


❑ TE content – IWGSC RefSeq v1

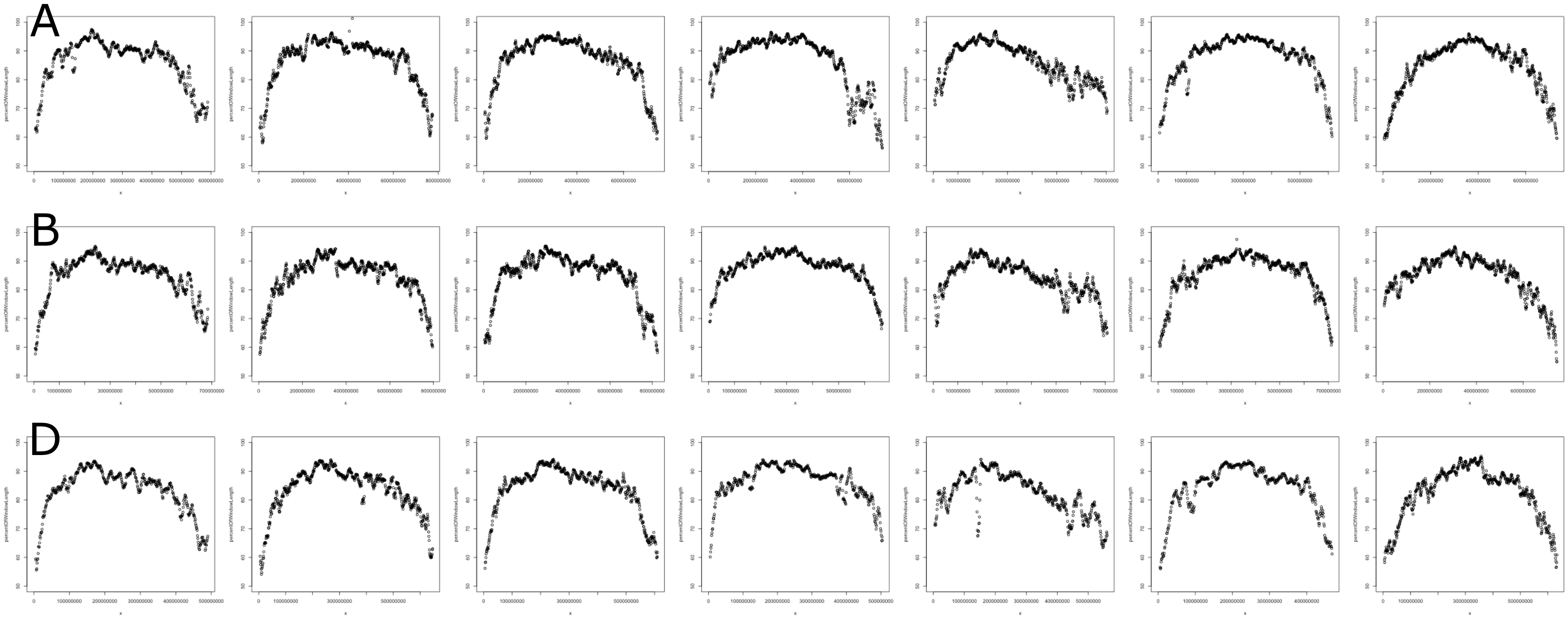
4.0M TEs (from 8.8M RepeatMasker matches)

14%

			A	B	D
All TEs		**84.8%**	86	85	83
ClassI	Gypsy	45.2%	51	47	41
	Copia	16.1%	17	16	16
	Others	3.1%			
	LINE	0.9%			
	SINE	0.0%			
ClassII	CACTA	**15.0%**	13	16	19
	Mutator	0.4%			
	Mariner	0.1%			
	Harbinger	0.1%			
	hAT	0.0%			
	Helitron	0.0%			
Unk	Unk	0.7%			


5 Gb
4 Gb
3 Gb
2 Gb
1 Gb
0 Gb
exons
non-TE
CACTA
Copia
Gypsy

○ TE distribution



○ Cereba+Quinta (centrom. gypsy families) distribution

A

B

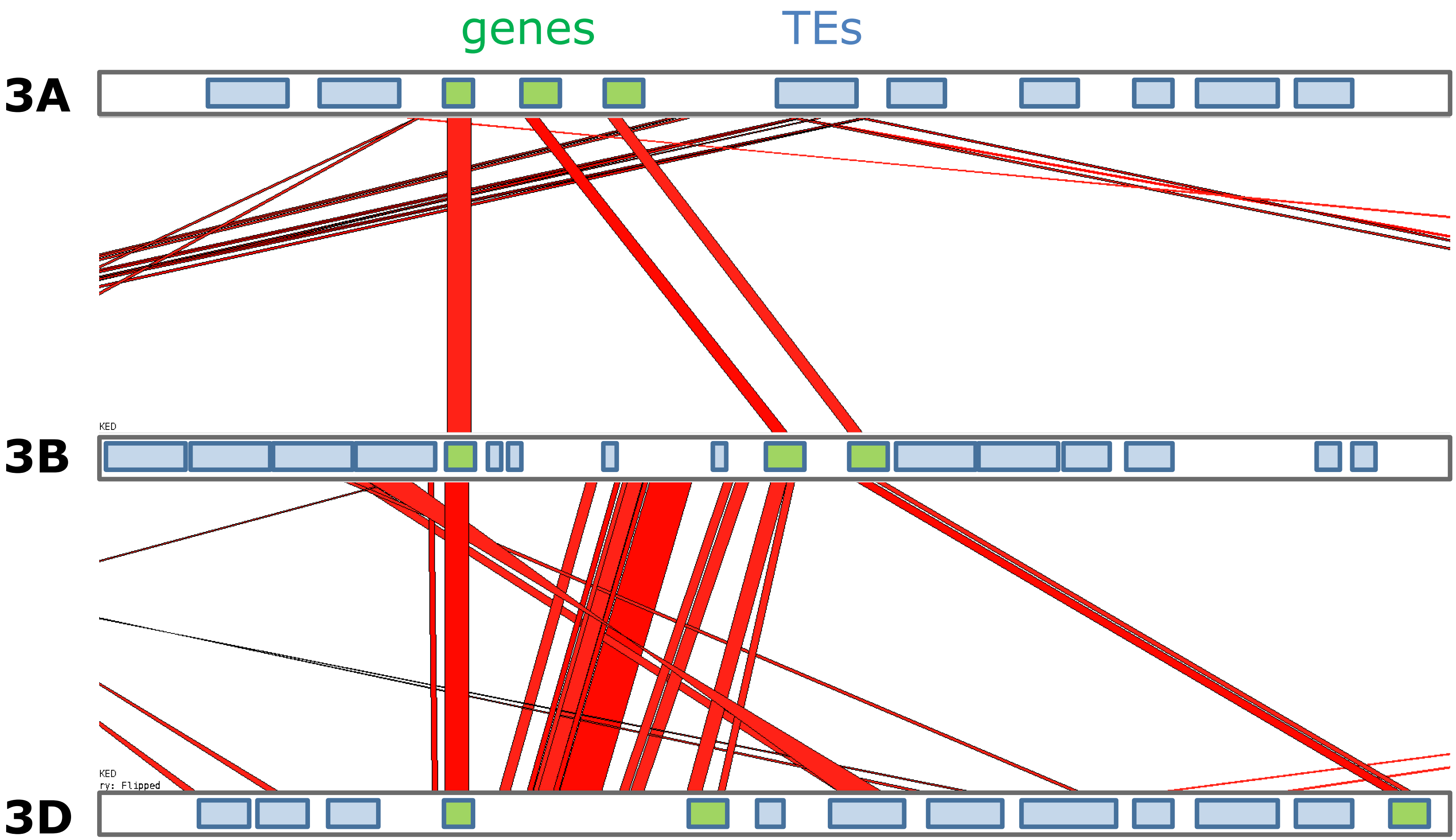
D

❑ TE: work in progress

- **Thomas Wicker**
  - explore the unannotated part of the genome
  - chromosome "niche" specificity

- **Heidrun Gunlach**
  - genome-wide characterization of the TE content and distribution

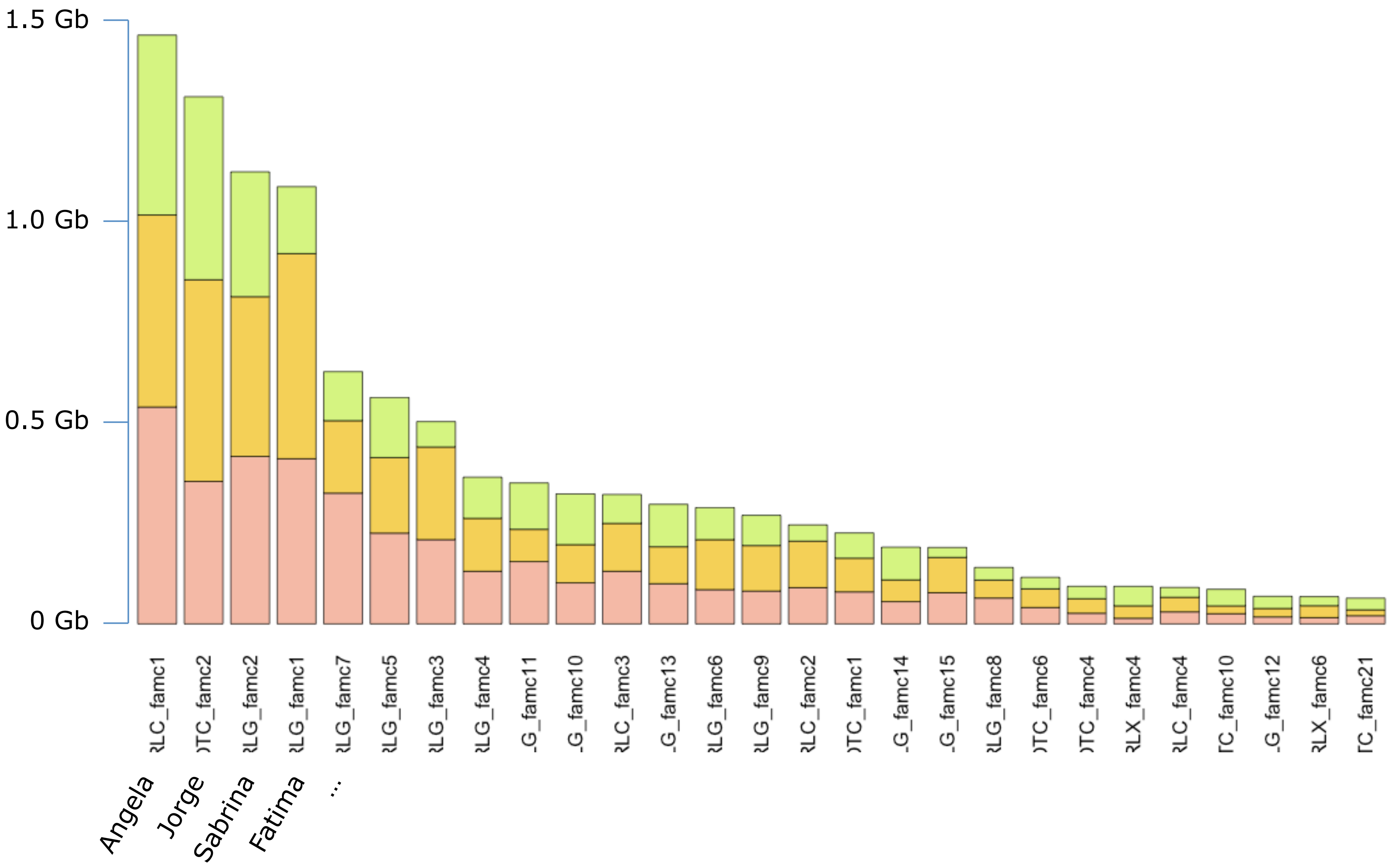
- **Frédéric Choulet**
  - evolutionary dynamics of wheat TEs
  - relationships TE/genes, especially CACTAs

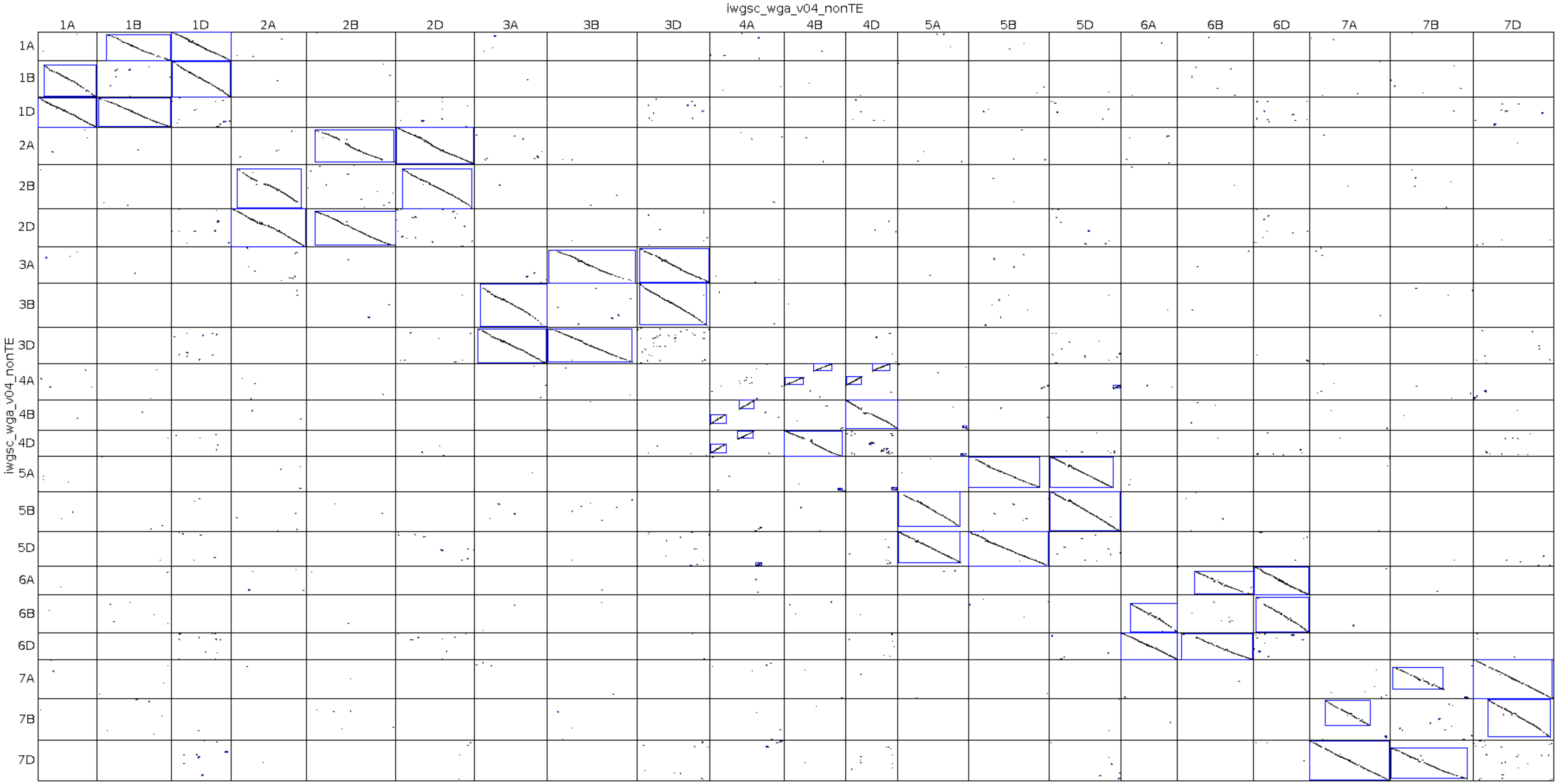

genes
TEs
3A
3B
3D
KED
KED
ry: Flipped

- What is the level of TE specificity in A-B-D?

1% (6/505) of TE families specific to 1 subgenome

➔ absence (almost) of subgenome specific TE families!!!


1.5 Gb
1.0 Gb
0.5 Gb
0 Gb
RLC_famc1
DTC_famc2
RLG_famc2
RLG_famc1
RLG_famc7
RLG_famc5
RLG_famc3
RLG_famc4
_G_famc11
_G_famc10
RLC_famc3
_G_famc13
RLG_famc6
RLG_famc9
RLC_famc2
DTC_famc1
_G_famc14
_G_famc15
RLG_famc8
DTC_famc6
DTC_famc4
RLX_famc4
RLC_famc4
TC_famc10
_G_famc12
RLX_famc6
TC_famc21
Angela
Jorge
Sabrina
Fatima
...


iwgsc_wga_v04_nonTE
1A
1B
1D
2A
2B
2D
3A
3B
3D
4A
4B
4D
5A
5B
5D
6A
6B
6D
7A
7B
7D
iwgsc_wga_v04_nonTE
1A
1B
1D
2A
2B
2D
3A
3B
3D
4A
4B
4D
5A
5B
5D
6A
6B
6D
7A
7B
7D

Acknowledgments

IWGSC-WGA working group

INRA GDEC

- Hélène Rimbert
- Philippe Leroy
- Romain De Oliveira
- Etienne Paux